AfriMMT-EA: Multi-domain Machine Translation for Low-Resource East African Languages

Anonymous EMNLP submission

Abstract

Despite remarkable progress in multilingual machine translation (MT), the majority of African—especially East African—languages remain significantly underrepresented both in benchmark datasets and state-of-the-art (SOTA) MT models. This persistent exclusion from mainstream technologies not only limits equitable access, but constrains the development of tools that accurately reflect the region's linguistic and cultural diversity. Recent advances in open-source large language models have demonstrated strong multilingual MT capabilities through data-efficient adaptation strategies. However, little work has explored their potential for low-resource African languages. We introduce AfriMMT-EA, the first highly multilingual benchmark and MT dataset for East African languages. Our datasets comprise 54 local languages across five East African countries. We used these data to fine-tune two multilingual versions of Gemma-3. We show improved translations with these models over a larger off-the-shelf baseline. We release our data and models, to advance MT for these low-resource languages and their communities.

1 Introduction

The pursuit of eliminating language barriers worldwide has positioned machine translation (MT) at the forefront of artificial intelligence research. The past few years have seen significant advancements in multilingual machine translation (MMT), largely driven by the growing capabilities of large language models (LLMs) such as GPT-4 (Hurst et al., 2024; Achiam et al., 2023; OpenAI, 2023), o1 (Jaech et al., 2024) LLaMA (Touvron et al., 2023; Dubey et al., 2024), and BLOOM (Le Scao et al., 2023), as well as the increased availability of multilingual training data for low-resource languages (LRLs) (Adelani et al., 2022b; Etori et al., 2025).

Earlier multilingual pretrained open-source models such as XLM-R (Conneau, 2019), mBART-50 (Tang et al., 2020), mBERT (Devlin, 2018), BLOOM (Le Scao et al., 2023), and NLLB-200 (Costa-Jussà et al., 2022) have demonstrated strong performance on multilingual tasks and are widely used as backbones for fine-tuning in low-resource MT research (Alabi et al., 2022a; Hedderich et al., 2020). However, these models were pretrained on corpora primarily sourced from Wikipedia and large-scale web crawls, with limited East African language coverage, and are increasingly outdated amid rapid open LLM advancements. Available MT corpora and resources lack support for a large number of LRLs.

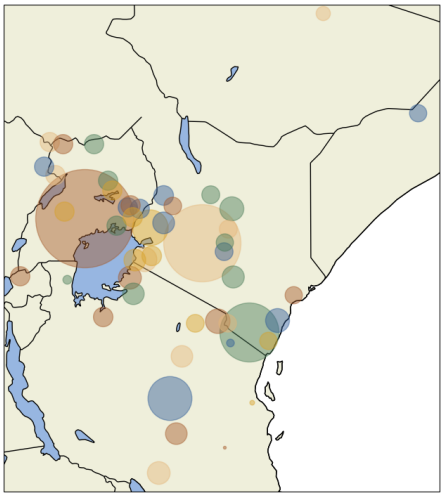

Figure 1: Visualization of the languages represented in our dataset. The size of each marker corresponds to the size of each language's bitext. Each language's geographical location is from Glottolog (Hammarström et al., 2023).

More recent open models, such as LLaMA3 (Touvron et al., 2023; Dubey et al., 2024), Mistral (Chaplot, 2023), Qwen3 (Yang et al., 2025), and Gemma (Team et al., 2025), have exhibited remarkable capabilities across various multilingual translation tasks and are being increasingly adapted for translation and cross-lingual applications (Guo et al., 2024; Li et al., 2024). However, the benefits of these advancements remain unevenly distributed

across the globe (Emezue and Dossou, 2022; Orife et al., 2020a; Adebara et al., 2023).

Models that power today's MT are trained mostly on data in high-resource languages such as English, French, German, or Russian, which have long enjoyed institutional investments and data abundance (Costa-Jussà et al., 2022). This leaves a massive gap for East African languages and potentially limits access to powerful technology for billions of people from the East African community (Akera et al., 2022a; Martinus and Abbott, 2019).

To address this gap, we present AfriMMT-EA, the first large-scale, multilingual, multi-domain MT datasets for East African languages. The dataset spans 54 local languages across Kenya, Uganda, Tanzania, Rwanda, and Ethiopia, comprising over 4M tokens and 714,490 parallel sentence pairs. Almost all Tanzanian languages and several Kenyan languages—including Hehe **(endangered)** and Chonyi **(extinct)**—are represented in MT research for the **first time**. Our dataset includes parallel corpora. We adopted a one-way translation approach; given that a few of our datasets had an English translation pair, while many more were translated to the local language, Swahili. Our datasets are carefully curated by humans (native speakers) to ensure localized and culturally appropriate translations. More specifically, we make the following contributions:

- We present a large dataset comprising 54 local East African languages, curated from diverse sources and domains and human-translated by native speakers. We detailed our specific contribution by tokens and translation pairs, alongside the multi-domain covered in table 4

- We introduce 11 Kenyan and 12 Tanzanian local languages into the mainstream of MT research, including coverage of one (1) endangered language , one (1) extinct languages and one (1) language [1]with emerging digital support for the first time in NLP.

- We demonstrate performance of modern MT models fine-tuned on our data across the various languages, highlighting better performance trends than a larger off-the-shelf model

[1]Language classification and vitality information were obtained from https://www.ethnologue.com/.

2 Related Work

2.1 Multilingual MT for African languages

Historically, MT research has focused on high-resource languages (Reid et al., 2021a; Robinson et al., 2024; Workshop et al., 2022), with much success in building performant systems. But African languages continue to be underrepresented, due to scarcity of quality training and evaluation data (Adelani et al., 2022a; Orife et al., 2020a; Ojo et al., 2023), leaving many African languages without quality MT (Akera et al., 2022a). There is a growing need for more inclusive datasets that capture linguistic diversity and support downstream tasks (Nekoto et al., 2020; Reid et al., 2021b). A robust body of work has attempted to address this underrepresentation of African languages. For instance, Google Translate,[2] a widely used MT interface, now supports five East African languages: Swahili, Luo, Oromo, Afar, and Somali. The Masakhane-MT project [3] (Orife et al., 2020b) includes MT datasets for several East African languages, including Swahili, Kikuyu, Luo, and Kamba. The primary source of data for these is the JW300 corpus (Agic and Vulic, 2019), and Swahili from the SAWA corpus (De Pauw et al., 2009). Swahili is the most commonly supported East African language in MT research. However, no prior MT work has explicitly targeted Swahili's Tanzania dialects. Recent advancements in MT LLMs (Achiam et al., 2023; Team et al., 2023; Cui et al., 2025) have expanded quality MT to more languages, but still neglect East African languages like Sukuma, Hehe, Teso, and Pokot.

2.2 Dataset creation

An MT system is only as good as the data it is trained on (Sloto et al., 2023). Parallel corpora form the backbone of neural MT training, providing the aligned bilingual data essential for learning high-quality translation mappings (Sánchez-Martínez et al., 2020; Brown et al., 1990). Nonetheless, the acquisition of high-quality parallel corpora, or bitexts, remains one of the most pressing challenges in MT research (Guo et al., 2018). Prior studies (Khayrallah and Koehn, 2018) have demonstrated that the presence of noise in parallel corpora—such as misaligned sentences, untranslated segments, or incorrect translations—can significantly degrade the performance of neural MT mod-

[2]https://translate.google.com
[3]https://www.masakhane.io/

els.[4] Currently, there are two main approaches to parallel data curation: (1) document and sentence alignment, and (2) comparable corpora methods (Sloto et al., 2023). We mostly utilize document alignment in this work. Recent efforts to improve the translation capabilities of LLMs have focused on leveraging multilingual data through continual pretraining of open-source models. For instance, Xu et al. (2023) conducted continual pretraining using only monolingual corpora, while Guo et al. (2024) extended this by incorporating parallel data after monolingual pretraining. Alves et al. (2024) proposed a mixed approach, utilizing a multilingual dataset composed of approximately two-thirds monolingual and one-third parallel data. Despite these advancements, the most effective strategy for combining monolingual and parallel corpora for MMT particularly for African languages, remains an open research question.

2.3 Evaluation benchmarks and metrics

Parallel corpora remain the gold standard for training and evaluating MT systems (Locke and Latham, 2002). Unfortunately, most LRLs lack digitized parallel data (Chimoto and Bassett, 2022), and the lack of standardized benchmarks and evaluation frameworks has been a barrier to LRL progress (Reid et al., 2021b; Duh et al., 2020). Recent efforts have emphasized the importance of participatory research in creating MT datasets (Kondoro, 2024; Babirye et al., 2022; Zawuya et al., 2025). But even large multilingual evaluations (e.g., MEGA (Ahuja et al., 2023) typically cover only a few African languages. This limits our ability to assess LLM performance across Africa's linguistic diversity and motivates benchmark efforts such as AFROBENCH (Ojo et al., 2023), which targeted multi-task evaluation across 64 African languages; or Toucan [5] (Elmadany et al., 2024a; Reid et al., 2021b), an Afrocentric many-to-many MT model that supports 156 African language pairs. For evaluation, Elmadany et al. (2024b) also developed AfroLingu-MT, a benchmark tailored for African MT. FLORES-101 (Goyal et al., 2022), FLORES200 (Team et al., 2022), Lafand-MT(Adelani et al., 2022b), and SALT (Akera et al., 2022b) also support some African languages. Other Afircan MT evaluation tools include AFRICOMET (Wang et al., 2023), a set of COMET-based[6] evaluation metrics tailored for African languages, and AfroXLM-R (Alabi et al., 2022b). Our work builds upon prior publications by standardizing evaluation sets for 54 languages, specifically from East Africa.

3 Methodology and Dataset

In this section, we detail the construction of our AfriMMT-EA dataset. Our dataset curation pipeline was based on the need and importance of enhancing the quality of datasets for multilingual language modeling (Oladipo et al., 2023). We conducted a rigorous examination and curation of existing data sources to ensure high data quality and linguistic relevance. This process prioritized the accuracy of aligned parallel data, essential for developing robust low-resource MT systems.

3.1 Language selection criteria

We selected data for East African languages spoken in Kenya, Uganda, Tanzania, Ethiopia, and Rwanda, based on two key considerations. First, two of the authors are native speakers from Kenya and Tanzania, offering valuable linguistic and cultural familiarity that informed language-specific decisions throughout the project. Second, as discussed in Section §2, the processes of LRL data acquisition, annotation, and sentence alignment are resource-intensive and time-consuming, and training data-hungry models on these data incurs computational and financial costs. Hence, our language selection was driven by dataset availability, rather than an attempt to achieve comprehensive regional coverage. Individual gatherers used best judgment collecting data sources for each language.

3.2 Data sources

Our dataset compilation draws on a range of multilingual and culturally grounded sources to ensure coverage across linguistic, cultural, and domain diversity. We collected data from the following countries and sources:

Kenya represents the largest contribution of the AfriMMT-EA, featuring over 25 languages. These resources provide extensive coverage across diverse domains such as religion, folklore, oral literature, education, governance, and health. The subset includes both widely spoken languages (e.g.,

[4] https://huggingface.co/spaces/evaluate-metric/bleu

[5] https://github.com/UBC-NLP/Toucan

[6] https://unbabel.github.io/COMET/html/index.html

Swahili, *Kikuyu*, *Luo*) and endangered or newly digitized languages (e.g., *Duruma*, *Taveta*, *Chonyi*) [7]. We included Somali language here. The Kenyan subset contributes over 270K translation pairs and nearly 8M tokens.

Most of the ***Ugandan*** language data originates from religious texts, particularly the Bible, supplemented by mixed-domain corpora covering education and general communication. Ugandan languages are digitally well-supported and classified as *ascending* in vitality (International, 2025) such as *Acholi*, *Alur*, *Luganda*, and *Soga*—while *Kebu* and *Kumam* are emerging languages with growing digital presence.

The ***Tanzanian*** languages data include religious, lexical, and crowdsourced corpora for languages such as *Sukuma*, *Makonde*, *Kihehe*, *Kichaga*, and *Zaramo*. The majority of these languages have not previously been incorporated into NLP datasets or resources.

Ethiopia contributes a small percentage of the data, with only two languages: *Oromo* and *Tigrinya*. The data are sourced from already published resources consists mixed domains such as education, governance, and health [8] *Oromo* is digitally supported and considered stable and ascending, while *Tigrinya* is institutionally supported and vital (International, 2025). Given different language family, Ethiopian languages were not our primary focus.

The *Kinyarwanda* datasets from ***Rwanda*** is primarily sourced from the Bible. It is also institutionally supported and classified as a vital language due to its strong digital presence (International, 2025).However, due to the single domain and language, Rwanda was not the primary focus in this work.

3.3 Data collection

Our dataset collection methodology is divided into three main steps: *searching and gathering*, *extraction* and *alignment*. We focused our initial data collection on **Kenya**, **Uganda**, and **Tanzania**, the foundational member states of the East African Community (EAC). These countries share **English** and **Swahili** with several native languages, supporting our one-way setup with limited English-to-local but ample Swahili-to-local data. Swahili serves as a pivot language for cross-lingual transfer and model alignment. To enhance regional diversity, we also sourced two languages from **Ethiopia** and one from **Rwanda**. In the future, we plan to expand our collection to more East African languages.

For each of the 54 target languages, we followed a structured data collection procedure. **(i)** First we found existing resources by querying research databases and search engines. Specifically, we searched for the language name in the *ACL Anthology*[9] and used combinations such as "`[language name]` machine translation", "`[language name]` NLP", "`[language name]` translation", and "`[language name]` African MT" on Google Scholar.[10] We expanded our search for parallel corpora using the Lanfrica database[11] and general Google search engine.[12] Given the large number of query results, we focused on the most relevant entries. **(ii)** Second, we searched online for books offering translations or supplementary language resources, including the Internet Archive.[13] **(iii)** Third, we contacted researchers and community members from the languages' communities to identify potential data sources such as Bible Translation & Literacy.[14] While response rates were low, these contacts enabled us to obtain some otherwise unavailable datasets. **(iv)** Fourth, we engaged native speakers by posting calls for contributions to translation tasks on social media such as Facebook [15] and LinkedIn. [16] This yielded some additional data, though many participants were unable to complete the translations fully. **(v)** Fifth, we established a web-based platform, *Argilla*, that enables East Africans to voluntarily contribute to MT data for their own languages[17]. A screenshot of the platform interface is in Figure 4.

Data Extraction: After data gathering, we categorized the resources into seven types: **(i)** previously published parallel datasets, **(ii)** web sources with pre-aligned parallel sentences, **(iii)** web sources containing full-text articles accompanied by translations, **(iv)** PDFs with aligned parallel sentences, **(v)** human-translated data generated via the Argilla platform or native speaker contributors, **(vi)** Bible parallel translations, and **(vii)** other PDF sources. Previously published parallel datasets

[7] https://www.ethnologue.com/
[8] https://github.com/AAUThematic4LT/Parallel-Corpora-for-Ethiopian-Languages
[9] https://aclanthology.org
[10] https://scholar.google.com
[11] https://lanfrica.com/
[12] https://www.google.com
[13] https://archive.org/
[14] https://www.btlkenya.org/
[15] https://www.facebook.com
[16] https://www.linkedin.com
[17] https://data-curator.sartify.com/sign-in

were combined at the beginning of the extraction process, including single-language resources such as KenTrans (Wanzare et al., 2022), Kencorpus (Wanjawa et al., 2022a), Lumasaba (Nabende et al., 2023), (Mbogho et al., 2025) and (Wanjawa et al., 2022b) . We customized extraction procedures for PDFs, websites, and other sources, leveraging *Gemini 2.5 Flash* (Comanici et al., 2025), with prompt-driven workflows followed by human-in-the-loop quality assurance. (See Fig. 2.) The extraction and annotations team included two native speakers of a Kenyan language, three native speakers of Tanzanian languages, and one Nigerian collaborator familiar with Ugandan languages. Although this did not cover all included languages, their familiarity with related regional languages enabled accurate processing. The team also ensured quality control in the datasets.

Data Alignment: The alignment process was crucial for transforming raw bilingual data into structured parallel data suitable for our downstream MT task. For web and document-based sources containing unsegmented text and their translations, we employed LLM-assisted segmentation and alignment with human verification. First, we copied text from the target languages directly from the original source text pair into the *Gemini 2.5 Flash* (Comanici et al., 2025) user interface, prompting it to identify and align corresponding sentence boundaries across source and target languages. The model segmented text based on punctuation, paragraph boundaries, and translation cues such as repetition or discourse markers. Next came our human validation phase. Annotators manually inspected the aligned segments to correct mismatches, missing sentences, or partial alignments. After verification, all aligned sentence pairs were organized and cleaned using spreadsheet tools Excel and Google Sheets. The aligned datasets were stored in comma-separated values (CSV) format.

3.4 Data preprocessing pipeline

The **AfriMMT-EA** Preprocessing pipeline was developed to construct a high-quality parallel multilingual translation dataset spanning 54 unique languages from **Kenya** (15 languages), **Uganda** (21 languages), **Tanzania** (12 languages), **Rwanda** (1), and **Ethiopia** (2). The pipeline follows a systematic sequence of steps: **(i)** data ingestion from heterogeneous CSV sources with automatic encoding detection, **(ii)** comprehensive text cleaning to remove encoding artifacts, normalize whitespace and punctuation, and filter empty or invalid entries, **(iii)** de-duplication of exact translation pairs, **(iv)** tokenization and metadata enrichment, **(v)** concurrent processing for efficient handling of multiple languages, **(vi)** deterministic train-test splitting per language pair, **(vii)** quality assurance and statistical validation, and **(viii)** dataset export and publication, including a comprehensive README with statistics and language coverage for reproducibility.

Data Organization and Ingestion: The raw dataset was organized hierarchically by country and language, with each language pair stored in separate CSV files. We employed automatic encoding detection to handle files with varying character encodings *UTF-8, UTF-8-BOM, Latin-1, CP1252, ISO-8859-1*, ensuring proper text extraction regardless of original file format. This multi-encoding support was critical given the diverse sources and historical nature of some texts. Files with missing or unreadable content were logged for inspection.

Text Cleaning: To address common data quality issues in multilingual corpora, each sentence was processed in a multi-stage cleaning procedure. The cleaning process began with systematic removal of encoding artifacts/erroneous UTF-8 symbols (e.g. $\surd$, ¬, ó)), byte-order marks, non-breaking spaces, and zero-width character artifacts from file format conversions and digitization processes. Regular expressions were used to identify and remove patterns such as standalone accented characters, negation artifacts, and malformed Unicode sequences. Next, we normalized whitespace and punctuation, removed standalone accents, converted ASCII, and removed translation pairs shorter than 3 characters or consisting solely of punctuation marks. We counted tokens using the Gemma-3 tokenizer to track source and translation token counts for each pair, enabling analysis of translation length ratios and informing model training decisions.

Deduplication and Dataset Splitting: To prevent data leakage and ensure model generalization, we implemented rigorous de-duplication. We removed translation pairs with identical source and translation text, and logged duplicate entries for inspection. The de-duplication process was applied globally across all languages to eliminate cross-contamination between language pairs.

For dataset splitting, we employed a stratified approach that maintains the distribution of language pairs in both training and test sets. Each language

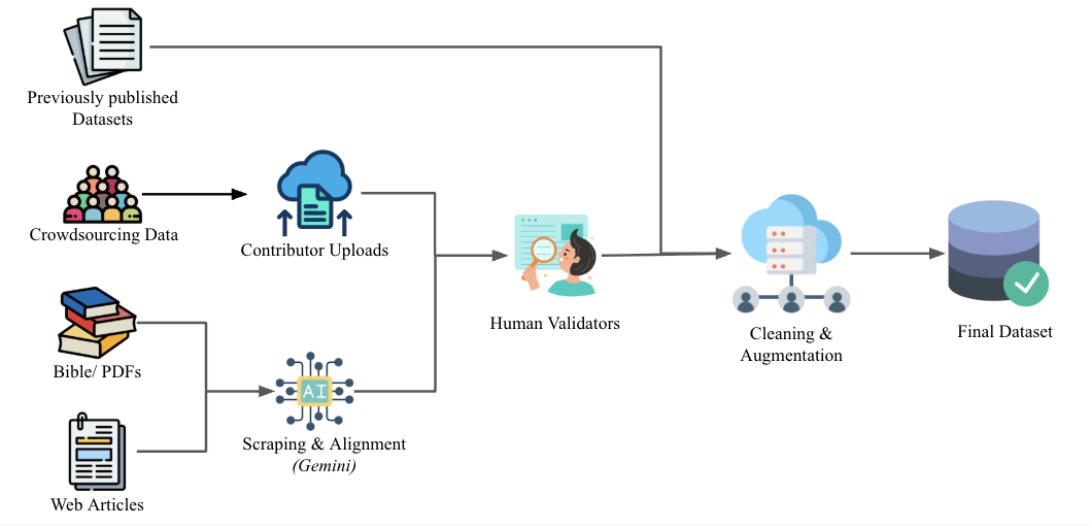

Figure 2: Data Collection Pipeline. Our framework integrates resource search, text extraction, and dual alignment (crowdsourced + Gemini-assisted) to produce validated East African multilingual data.

pair was independently shuffled using a fixed random seed (42) for reproducibility, then split using an 80/20 ratio. A similar approach that maximizes data coverages was employed for languages with fewer pairs. This stratified splitting ensures that every language pair is adequately represented in both splits while maintaining evaluation integrity.

The final dataset was augmented with metadata including **source language**, **target language**, **country of origin**, **token counts**, **processing timestamp**, and **processor version** for full traceability. Statistical validation confirmed zero remaining duplicates post-processing and verified the 80/20 split ratio across all language pairs.

Data distribution Table 1 presents the country-level data distribution across **AfriMMT-EA**. The dataset covers a total of 54 languages, with **Kenya** contributing the largest proportion. Kenya and Uganda together account for over 80% of the data. Table 3 details the distribution of translation pairs, token counts, and language representation, while Table 2 provides information about each language in AfriMM-EA.

4 Modeling Experiments

To evaluate the dataset's utility for MT tasks involving East African languages, we fine-tuned a multilingual translation model on our dataset and evaluated its performance across various language pairs.

Country	Languages	Translation Pairs	Percentage
Kenya	24	351,396	48.8%
Uganda	16	256,133	35.5%
Tanzania	12	103,520	14.4%
Rwanda	1	6,246	0.9%
Ethiopia	2	3,441	0.5%

Table 1: **Geographic Distribution** of translation pairs by country. Kenya and Uganda dominate the dataset, together accounting for over 80% of the total data.

4.1 Experimental Setup

"We utilized a pre-trained language model as our base architecture and fine-tuned all model parameters during training. This strategy notably reduces the number of trainable parameters while preserving the model's expressive capability for the translation task.

The model was configured with a maximum sequence length of 2,048 tokens to accommodate longer translations. We utilized the Gemma-3 chat template to format training examples, structuring each instance as a three-turn conversation: a system message defining the translation task, a user message containing the source text with language pair specification, and an assistant response containing the target translation. The training objective was optimized to predict only the assistant's response, masking losses on the instruction portions using response-only training. Hyperparameter optimization was conducted using Weights & Biases sweeps to identify optimal training configurations. The training process used the AdamW 8-bit optimizer

ISO Code	Name	Family	Region	Speakers	Vitality	Digital support
swa	Swahili	Niger_Congo/Bantu	East and Central	1M -1B	Institutional	Vital
dug	Duruma	Niger-Congo/Bantu	Kenya	10K- 1M	Institutional	Ascending
sxb	Suba	Niger-Congo/Bantu	Kenya and Tanzania	10K- 1M	Stable	Ascending
dav	Taita	Niger-Congo/Bantu	Kenya	10K- 1M	Stable	Ascending
mer	Meru	Niger-Congo/Bantu	Kenya	1M- 2B	Institutional	Ascending
guz	Gusii	Niger-Congo/Bantu	Kenya and Tanzania	1M- 1B	Stable	Ascending
lsm	Saamia	Niger-Congo/Bantu	Kenya and Uganda	10K-1M	Institutional	Emerging
kam	Kamba	Niger-Congo/Bantu	Kenya	1M- 1B	Institutional	Ascending
teo	Teso	Nilo-Saharan	Kenya and Uganda .	1M- 1B	Institutional	Ascending
kln	Kalenjin	Nilo-Saharan	Kenya	1M- 1B	Stable	Ascending
ebu	Embu	Niger-Congo/Bantu	Kenya	10K-1M	Stable	Ascending
coh	Chonyi	Niger-Congo/Bantu	Coastal Kenya	10K-1M	Stable	Still
lwg	Luwanga	Niger-Congo/Bantu	Western Kenya	10K-1M	Stable	Emerging
kik	Kikuyu	Niger-Congo/Bantu	Central Kenya	1M -1B	Stable	Ascending
luo	Luo	Nilo-Saharan	Kenya and Tanzania	1M -1B	Stable	Ascending
mas	Maasai	Nilo-Saharan	Kenya and Tanzania	1M -1B	Stable	Ascending
gax	Borana	Afro-Asiatic	Kenya and Ethiopia	1M -1B	Stable	Ascending
thk	Tharaka	Niger-Congo/Bantu	Kenya	10K-1M	Institutional	Ascending
bxk	Bukusu	Niger-Congo/Bantu	Kenya and Uganda	1M -1B	Institutional	Ascending
rag	Maragoli	Niger–Congo/Bantu	Kenya	10K-1M	Institutional	Ascending
cuh	Chuka	Niger–Congo/Bantu	Kenya	10K-1M	Stable	Emerging
saq	Samburu	Nilo-Saharan	Kenya	10K-1M	Stable	Ascending
enb	Marakwet	Nilo-Saharan	Kenya	10K-1M	Stable	Ascending
tvs	Taveta	Niger–Congo/Bantu	Kenya	10K-1M	Stable	Emerging
nyf	Giriama	Niger–Congo/Bantu	Kenya	1M -1B	Institutional	Ascending
pkb	Pokomo	Niger–Congo/Bantu	Kenya	10K-1M	Institutional	Ascending
pko	Pokot	Nilo-Saharan	Kenya	1M -1B	Stable	Ascending
som	Somali	Afro-Asiatic	Kenya and Somalia	1M -1B	Institutional	Vital
gax	Oromo	Afro-Asiatic	Ethiopia	1M–1B	Stable	Ascending
tir	Tigrinya	Afro-Asiatic	Ethiopia	1M–1B	Institutional	Vital
kin	Kinyarwanda	Niger-Congo/Bantu	Rwanda	1M–1B	Institutional	Vital
bez	Bena	Niger-Congo	Central Tanzania	10K - 1M	Stable	Ascending
gog	Gogo	Niger-Congo	Central Tanzania	1M - 1B	Stable	Ascending
kde	Makonde	Niger-Congo	SE. Tanzania, No. Mozambique	1M - 1B	Stable	Ascending
lag	Langi	Niger-Congo	Tanzania	10K - 1M	Stable	Ascending
mas	Maasai	Nilo-Saharan	Tanzania, Kenya	1M - 1B	Stable	Ascending
jit	Jita	Niger-Congo	Tanzania	10K - 1M	Stable	Ascending
heh	Kihehe	Niger-Congo	Southern Tanzania	1M - 1B	Endangered	Ascending
old	Kichaga	Niger-Congo	Kilimanjaro, Tanzania	10K - 1M	Stable	Emerging
suk	Kisukuma	Niger-Congo	Northern Tanzania	1M - 1B	Stable	Ascending
hay	Kihaya	Niger-Congo/Bantu	Northwestern Tanzania	1M - 1B	Stable	Ascending
asu	Pare	Niger-Congo	Northeast Coast Tanzania	10K - 1M	Stable	Ascending
ach	Acholi	Nilo-Saharan	Uganda, South Sudan	2M	Institutional	Ascending
alz	Alur	Nilo-Saharan	DRC, Uganda	2.5M	Institutional	Ascending
luc	Aringa	Nilo-Saharan	Uganda	927,000.	Stable	Ascending
teo	Ateso	Nilo-Saharan	Uganda, Kenya	3.1M	Institutional	Ascending
gwr	Gwere	Niger-Congo	Uganda	876,000	Institutional	Ascending
adh	Jopadhola	Nilo-Saharan	Uganda	570,000	Stable	Ascending
keo	Kakwa	Nilo-Saharan	Uganda, DRC, South Sudan	591,000	Stable	Ascending
ndp	Kebu	Nilo-Saharan	Uganda	51,700	Stable	Emerging
kdi	Kumam	Nilo-Saharan	Uganda	350,000	Institutional	Emerging
laj	Lango	Nilo-Saharan	Uganda	2.7M	Institutional	Ascending
lgg	Lugbara	Nilo-Saharan	Uganda, DR Congo	1.2M	Institutional	Ascending
myx	Masaaba	Niger-Congo	Uganda	2.1M	Institutional	Ascending
nuj	Nyole	Niger-Congo	Uganda	633,000	Stable	Ascending
nyo	Nyoro	Niger-Congo	Uganda	1.2M	Institutional	Ascending
xog	Soga	Niger-Congo	Uganda	3.7M	Institutional	Ascending

Table 2: ISO Language Codes, Names, Language Families, Regions, Speaker Estimates, Vitality, and Digital Language Support, per Ethnologue (https://www.ethnologue.com).

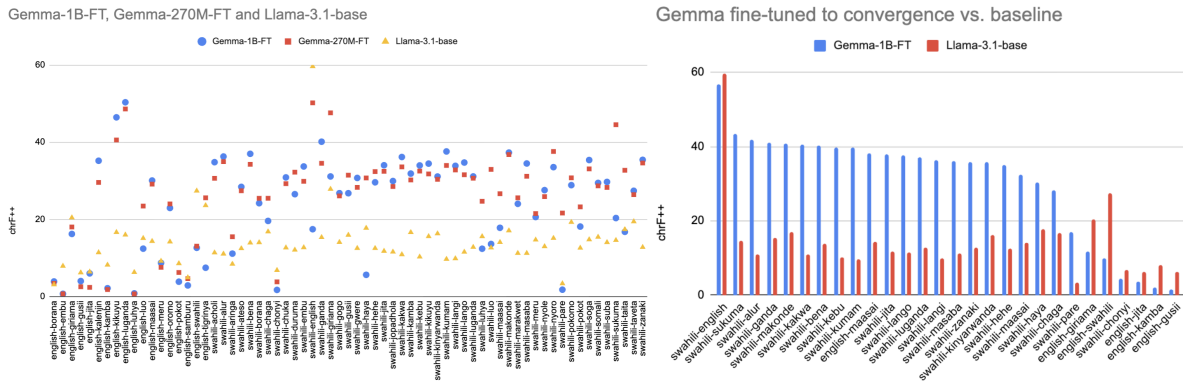

Figure 3: ***Left:*** comparison chrF++ scores for two fine-tuned models and a larger (7B-parameter) off-the-shelf Llama 3.1 baseline. Due to compute and time constraints as the dataset was finalized, these models were not trained to convergence. ***Right:*** ChrF++ for a Gemma model fine-tuned to convergence on a subset of the language pairs represented in AfriMM-EA, compared with corresponding scores for the same baseline Llama 3.1 7B-parameter model. "**FT**" indicates models fine-tuned on AfriMM-EA.

for memory efficiency, with batch sizes and gradient accumulation steps varied across experiments. We employed a linear learning rate schedule with 5 warmup steps. The learning rate, weight decay, number of training epochs, and effective batch size (product of per-device batch size and gradient accumulation steps) were treated as hyperparameters subject to optimization. All experiments used gradient checkpointing via the Unsloth framework (Han et al., 2023) to reduce memory consumption during training.

For evaluation, we assessed translation quality across all language pairs in the test set using three complementary metrics: BLEU (Papineni et al., 2001), chrF++ (Popović, 2017), and TER (Snover et al., 2006). Translations were generated using greedy decoding (temperature=0.0) with a maximum length of 128 tokens. We employed batch inference with a batch size of 256 to accelerate evaluation, processing the entire test set efficiently. The `sacrebleu` library (Post et al., 2025) with international tokenization was used for all metric computations. Results were aggregated by language pair, and overall performance was computed as weighted averages based on the number of samples per pair.

4.2 Results and discussion

We discuss insights from our trained model across various dimensions, including translation pair, model size, and overall performance.

We present the overall results across all 67 AfriMM-EA translation pairs in our dataset pair in Fig. 3. Using the chrF++ metrics from the results (Fig. 3 Right), we found that the MT models performed best for the translation pairs that translate English to popular East African languages including english→luganda (49.47) and english→kikuyu (43.40); however, the model performed worst for languages with lower resource translation pairs, which include english→embu and english→luhya where both model has an average chrF++ of 0.64 and 0.70 respectively

Across model sizes, we noticed that Gemini-3-270M-it had better performance with an average chrF++ of 26.35 as compared with the large Gemini-3-1B-it with an average chrF++ of 23.74, suggesting that smaller parameter-sized models might be better at the MT task for low-resource languages. We then selected a subset of 13 languages that have bilingual pairs in English and Swahili. Our results in table 5 (top rows) showed that the models are about twice as good in translating text in the languages Swahili and English. This is also in line with the availability of more Swahili paired datasets in this language in contrast to English. See table appendex. Specifically, both model achieved an average crF++ score of 27.6% for their Swahili translation in contrast to an average crF++ score of 27.6% for English paired translation. This result is not surprising as Swahili is the most spoken language across these regions.

4.3 Conclusion

In this work, we presented **AfriMMT-EA**, a multi-domain machine translation benchmark and dataset covering 54 low-resource East African languages to advance multilingual low-resource research.

Limitations

Like any study or data curation process, we acknowledge certain limitations in the current work. Firstly, this study is delimited to 57 languages from four East African countries: Tanzania, Uganda, Kenya, and Ethiopia. While the initial goal was to include additional East African countries, we were constrained by the availability of data and the resources needed to recruit more human curators. Consequently, we limited the scope of the study to the aforementioned countries, which share similar geographical, socio-cultural, and linguistic contexts. We believe, however, that this study outlines a roadmap for future work that could expand to include machine translation (MT) for more East African languages.

Secondly, the training of the MT model was constrained by the availability of computing power and GPU resources. As a result, our experiments were limited to fine-tuning two Gemini models, which did not fully converge. Future research will aim to include a broader range of models and explore their performance in more depth.

Additionally, our dataset collection includes publicly released datasets that are under the CC BY License. We have made every effort to properly attribute the sources of these datasets. The datasets also include contributions from human annotators, who are credited as authors in the publications.

Ethics Statement

Because East African languages have historically been underrepresented and marginalized in linguistic research and digital technologies, it is essential to approach their inclusion in multilingual NLP with cultural and ethical sensitivity. The linguistic landscape of East Africa is shaped by centuries of multilingual interaction, colonial influence, and sociolinguistic hierarchies that often privilege former colonial, such as English, over indigenous languages. Our work seeks to counter this imbalance by promoting equitable language representation while recognizing that technological interventions can have uneven social impacts.

We acknowledge that not all speakers of East African languages may view machine translation or language modeling as inherently beneficial. Some communities may be concerned about potential misuse of data, the commodification of their linguistic heritage, or the risk of cultural dilution through automated systems. Others may welcome these tools as enablers of educational access, digital inclusion, and language preservation. We believe this research has positive social impact by promoting linguistic inclusivity and supporting language technology development in low-resource communities. The texts included in our dataset reflect a wide range of cultural, religious, and historical viewpoints. Some older materials, such as religious or folklore texts, may contain ideas that are considered outdated or culturally sensitive in modern contexts. We do not endorse these views but retain such data to preserve linguistic and cultural authenticity. All collected materials were either publicly available or contributed with informed consent, and any potentially identifiable personal information was removed or anonymized during processing.

We recognize that our dataset may not fully capture the linguistic diversity or dialectal variation present within each language community, leading to potential biases or representational gaps. We encourage future work to report results across dialects, domains, and community varieties to better reflect linguistic reality. Furthermore, we discourage any use of our resources for surveillance, exclusionary policies, or applications that could marginalize vulnerable linguistic groups.

Finally, we acknowledge that AI assistance (e.g., ChatGPT, GEMINI) was used to support writing and documentation in this research. All outputs were reviewed and verified by the authors to ensure accuracy, cultural integrity, and compliance with the ACL Code of Ethics.

Acknowledgments

Anonymized for review.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2023. SERENGETI: Massively Multilingual Language Models for Africa. ArXiv:2212.10785 [cs].

David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Chinenye

Emezue, Colin Leong, Michael Beukman, Shamsuddeen Hassan Muhammad, Guyo Dub Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ayoade Ajibade, Tunde Oluwaseyi Ajayi, Yvonne Wambui Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Koffi Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022a. A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation. ArXiv:2205.02022 [cs].

David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, et al. 2022b. A few thousand translations go a long way! leveraging pre-trained models for african news translation. arXiv preprint arXiv:2205.02022.

Zeljko Agic and Ivan Vulic. 2019. Jw300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. Available at: https://scholar.google.com/scholar?cluster=8239399398272024687&hl=en&as_sdt=0,24.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Benjamin Akera, Jonathan Mukiibi, Lydia Sanyu Naggayi, Claire Babirye, Isaac Owomugisha, Solomon Nsumba, Joyce Nakatumba-Nabende, Engineer Bainomugisha, Ernest Mwebaze, and John Quinn. 2022a. Machine Translation For African Languages: Community Creation Of Datasets And Models In Uganda.

Benjamin Akera, Jonathan Mukiibi, Lydia Sanyu Naggayi, Claire Babirye, Isaac Owomugisha, Solomon Nsumba, Joyce Nakatumba-Nabende, Engineer Bainomugisha, Ernest Mwebaze, and John Quinn. 2022b. Machine translation for african languages: Community creation of datasets and models in uganda. In *3rd Workshop on African Natural Language Processing*.

Jesujoba O Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022a. Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning. *arXiv preprint arXiv:2204.06487*.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022b. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Claire Babirye, Joyce Nakatumba-Nabende, Andrew Katumba, Ronald Ogwang, Jeremy Tusubira Francis, Jonathan Mukiibi, Medadi Ssentanda, Lilian D. Wanzare, and Davis David. 2022. Building Text and Speech Datasets for Low Resourced Languages: A Case of Languages in East Africa.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Devendra Singh Chaplot. 2023. Albert q. jiang, alexandre sablayrolles, arthur mensch, chris bamford, devendra singh chaplot, diego de las casas, florian bressand, gianna lengyel, guillaume lample, lucile saulnier, lélio renard lavaud, marie-anne lachaux, pierre stock, teven le scao, thibaut lavril, thomas wang, timothée lacroix, william el sayed. *arXiv preprint arXiv:2310.06825*, 3.

Everlyn Asiko Chimoto and Bruce A Bassett. 2022. Very low resource sentence alignment: Luhya and swahili. *arXiv preprint arXiv:2211.00046*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. *arXiv preprint arXiv:2502.02481*.

Guy De Pauw, Peter Waiganjo Wagacha, and Gilles-Maurice de Schryver. 2009. The sawa corpus: a parallel corpus english-swahili. In *Proceedings of the First Workshop on Language Technologies for African Languages*, pages 9–16.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kevin Duh, Paul McNamee, Matt Post, and Brian Thompson. 2020. Benchmarking neural and statistical machine translation on low-resource African languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2667–2675.

AbdelRahim Elmadany, Ife Adebara, and Muhammad Abdul-Mageed. 2024a. Toucan: Many-to-many translation for 150 african language pairs. *arXiv preprint arXiv:2407.04796*.

AbdelRahim Elmadany, Ife Adebara, and Muhammad Abdul-Mageed. 2024b. Toucan: Many-to-Many Translation for 150 African Language Pairs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13189–13206, Bangkok, Thailand. Association for Computational Linguistics.

Chris C. Emezue and Bonaventure F. P. Dossou. 2022. MMTAfrica: Multilingual Machine Translation for African Languages. ArXiv:2204.04306 [cs].

Naome A Etori, Kevin Lu, Randu Karisa, and Arturs Kanepajs. 2025. Lag-mmlu: Benchmarking frontier llm understanding in latvian and giriama. *arXiv preprint arXiv:2503.11911*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Jiaxin Guo, Hao Yang, Zongyao Li, Daimeng Wei, Hengchao Shang, and Xiaoyu Chen. 2024. A novel paradigm boosting translation capabilities of large language models. *arXiv preprint arXiv:2403.11430*.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, et al. 2018. Effective parallel corpus mining using bilingual sentence embeddings. *arXiv preprint arXiv:1807.11906*.

Harald Hammarström, Sebastian Nordhoff, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2023. Glottolog. Online database.

Daniel Han, Michael Han, and Unsloth team. 2023. Unsloth.

Michael A Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow. 2020. Transfer learning and distant supervision for multilingual transformer models: A study on african languages. *arXiv preprint arXiv:2010.03179*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

SIL International. 2025. Ethnologue: Languages of the world. Accessed: 2025-10-06.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. *arXiv preprint arXiv:1805.12282*.

Alfred Malengo Kondoro. 2024. SwaQuAD-24: QA Benchmark Dataset in Swahili. ArXiv:2410.14289 [cs].

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagnè, Alexandra Sasha Luccioni, François Yvon, and Matthias Gallé. 2023. Bloom: A 176b-parameter open-access multilingual language model. Preprint. Available at `https://inria.hal.science/hal-03850124/`.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *Transactions of the Association for Computational Linguistics*, 12:576–592.

Edwin A. Locke and Gary P. Latham. 2002. Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, 57(9):705–717.

Laura Martinus and Jade Z. Abbott. 2019. A Focus on Neural Machine Translation for African Languages. ArXiv:1906.05685 [cs].

Audrey Mbogho, Quin Awuor, Andrew Kipkebut, Lilian Wanzare, and Vivian Oloo. 2025. Building low-resource african language corpora: A case study of kidawida, kalenjin and dholuo. arXiv preprint arXiv:2501.11003.

Peter Nabende, Naomi Muzaki, Claire Babirye, Jonathan Mukiibi, Jeremy Tusubira, Joyce Nakatumba-Nabende, and Andrew Katumba. 2023. Lumasaba Monolingual Corpus.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo, Salomey Osei, Sackey Freshia, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa, Mofe Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Jane Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkabir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Espoir Murhabazi, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Emezue, Bonaventure Dossou, Blessing Sibanda, Blessing Itoro Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. ArXiv:2010.02353 [cs].

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2023. Afrobench: How good are large language models on african languages. In *18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1963–1978.

Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Abraham Owodunni, Odunayo Ogundepo, David Adelani, and Jimmy Lin. 2023. Better quality pre-training data and t5 models for african languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 158–168.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, Kevin Degila, Herman Kamper, Bonaventure Dossou, Chris Emezue, Kelechi Ogueji, and Abdallah Bashir. 2020a. Masakhane – Machine Translation For Africa. ArXiv:2003.11529 [cs].

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, et al. 2020b. Masakhane–machine translation for africa. *arXiv preprint arXiv:2003.11529*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 311, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post, Martin Popel, and Ozan Caglayan. 2025. Sacrebleu. Version 2.5.1.

Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. 2021a. AfroMT: Pretraining strategies and reproducible benchmarks for translation of 8 african languages. arXiv preprint arXiv:2109.04715.

Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. 2021b. AfroMT: Pretraining Strategies and Reproducible Benchmarks for Translation of 8 African Languages. ArXiv:2109.04715 [cs].

Nathaniel R Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Bizon Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome A Etori, et al. 2024. Krey\ol-mt: Building mt for latin american, caribbean and colonial african creole languages. *arXiv preprint arXiv:2405.05376*.

Felipe Sánchez-Martínez, Víctor M Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, Mikel L Forcada, Miquel Espla-Gomis, Andrew Secker, Susie Coleman, and Julie Wall. 2020. An english-swahili parallel corpus and its use for neural machine translation in the news domain. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 299–308.

Steve Sloto, Brian Thompson, Huda Khayrallah, Tobias Domhan, Thamme Gowda, and Philipp Koehn. 2023. Findings of the wmt 2023 shared task on parallel data curation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 95–102.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane

Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

NLLB Team, Marta R. Costa-jussà, James Cross andOnur Çelebiand Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang anf Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, et al. 2023. Afrimte and africomet: Enhancing comet to embrace under-resourced african languages. *arXiv preprint arXiv:2311.09828*.

Barack Wanjawa, Lilian D.A. Wanzare, Florence Indede, Owen McOnyango, Edward Ombui, and Lawrence Muchemi. 2022a. Kencorpus: Kenyan Languages Corpus.

Barack Wanjawa, Lilian D.A. Wanzare, Florence Indede, Owen McOnyango, Edward Ombui, and Lawrence Muchemi. 2022b. Kencorpus: Kenyan languages corpus.

Lilian D.A Wanzare, Florence Indede, Owen McOnyango, Edward Ombui, Barack Wanjawa, and Lawrence Muchemi. 2022. KenTrans: A Parallel Corpora for Swahili and local Kenyan Languages.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Chaddy Anthony Zawuya, Alfred Malengo Kondoro, Diana Rwegasira, and Juma H. Lungo. 2025. Maneno yetu: Dynamic corpus construction and pretraining for swahili nlp. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, New York, NY, USA. Association for Computing Machinery.

A Argilla Web-based Translation Interface

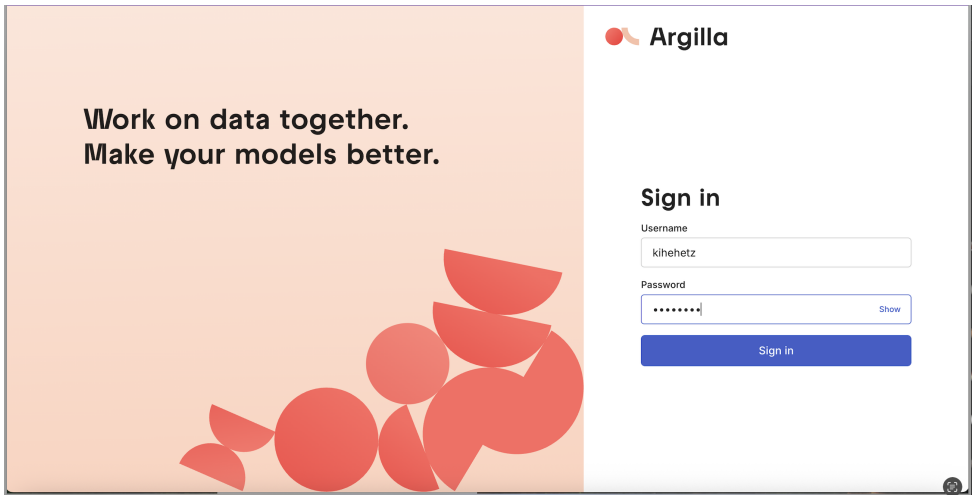

Figure 4: Argilla web-based translation interface. Each contributor is provided with a unique username and password for the specific language they wish to contribute to, allowing controlled access to submit translation data.

B Instruction to Volunteers

C Data Format

Data Format: Our aligned datasets were stored in CSV format using a schema shown in Table **??** below. Each record represents a single translation pair with metadata describing its source, processing stage, and origin.

D Language Distribution

Our datasets comprises translation pairs across 52 East African languages from five countries: Uganda, Kenya, Tanzania, Rwanda, and Ethiopia. Table 3 presents the distribution of translation pairs, token counts, and language representation across the dataset. The distribution of languages in the dataset is uneven. Luganda (Uganda) dominates, accounting for 22.7% of the corpus with 162,446 translation pairs and over 8 million tokens. Kikuyu (Kenya) is the second most represented language at 14.0%, followed by Taita (Kenya) at 8.2% and English (Kenya) at 7.6%. Collectively, these four languages constitute more than half of the dataset.

The majority of languages (37 out of 52) represent less than 1% of the dataset each, highlighting the long-tail distribution typical of low-resource language collections. Kenya contributes the most languages to the dataset (21 languages), followed by Uganda (17 languages) and Tanzania (11 languages). Rwanda and Ethiopia contribute one and two languages respectively.

Token counts vary significantly across languages, with source tokens ranging from 408 (Tigrinya) to 3,079,301 (Luganda), and translation tokens ranging from 90 (Tigrinya) to 5,061,186 (Luganda). This variation reflects both the different dataset sizes and the morphological characteristics of each language.

Table 3: Language Statistics by Country

Language	Country	Translation Pairs	Source Tokens	Translation Tokens	Total Tokens	Percentage of language
Luganda	Uganda	162,446	3,079,301	5,061,186	8,140,487	22.7%
Kikuyu	Kenya	100,195	2,141,717	3,510,370	5,652,087	14.0%
Taita	Kenya	58,537	809,231	888,257	1,697,488	8.2%
English	Kenya	54,221	1,020,199	739,610	1,759,809	7.6%
Gogo	Tanzania	32,371	1,456,656	1,865,915	3,322,571	4.5%
Luhya	Kenya	21,560	317,414	339,491	656,905	3.0%
Chaga	Tanzania	10,106	351,653	393,654	745,307	1.4%
Luo	Kenya	10,038	362,766	356,761	719,527	1.4%
Borana	Kenya	9,697	247,267	264,151	511,418	1.4%
Giriama	Kenya	9,629	327,412	376,830	704,242	1.3%
Makonde	Tanzania	9,111	395,038	516,338	911,376	1.3%
Jita	Tanzania	8,820	333,545	361,195	694,740	1.2%
Bena	Tanzania	8,674	334,552	480,999	815,551	1.2%
Kamba	Kenya	8,210	212,480	220,388	432,868	1.1%
Kalenjin	Kenya	7,969	84,952	135,554	220,506	1.1%
Suba	Kenya	7,965	361,159	437,423	798,582	1.1%
Hehe	Tanzania	7,849	325,353	365,460	690,813	1.1%
Langi	Tanzania	7,849	322,965	409,292	732,257	1.1%
Zanaki	Tanzania	7,813	321,367	485,813	807,180	1.1%
Pokot	Kenya	6,533	234,343	217,501	451,844	0.9%
Gusii	Kenya	6,494	269,601	261,580	531,181	0.9%
Sukuma	Tanzania	6,394	192,310	227,498	419,808	0.9%
Masaba	Uganda	6,246	231,741	328,161	559,902	0.9%
Acholi	Uganda	6,246	231,740	242,244	473,984	0.9%
Alur	Uganda	6,246	231,740	231,136	462,876	0.9%
Ateso	Uganda	6,246	231,740	280,299	512,039	0.9%
Kakwa	Uganda	6,246	231,740	289,423	521,163	0.9%
Gwere	Uganda	6,246	231,740	340,491	572,231	0.9%
Jopadhola	Uganda	6,246	231,740	247,727	479,467	0.9%
Lango	Uganda	6,246	231,741	237,077	468,818	0.9%
Kebu	Uganda	6,246	231,740	265,804	497,544	0.9%
Kinyarwanda	Rwanda	6,246	231,740	235,655	467,395	0.9%
Nyole	Uganda	6,246	231,740	294,609	526,349	0.9%
Nyoro	Uganda	6,246	231,740	246,490	478,230	0.9%
Soga	Uganda	6,246	231,740	267,777	499,517	0.9%
Kumam	Uganda	6,246	231,740	265,015	496,755	0.9%
Aringa	Uganda	6,243	231,605	229,863	461,468	0.9%
Meru	Kenya	5,420	207,053	206,282	413,335	0.8%
Samburu	Kenya	5,411	47,140	45,595	92,735	0.8%
Embu	Kenya	5,349	228,901	226,948	455,849	0.7%
Maasai	Tanzania	5,327	154,089	190,289	344,378	0.7%
Taveta	Kenya	5,249	228,860	186,804	415,664	0.7%
Marakwet	Kenya	5,249	228,624	266,270	494,894	0.7%
Chuka	Kenya	5,030	205,331	195,227	400,558	0.7%
Duruma	Kenya	5,011	197,413	248,717	446,130	0.7%
Somali	Kenya	5,011	197,413	235,446	432,859	0.7%
Pokomo	Kenya	5,010	197,338	246,456	443,794	0.7%
Oromo	Ethiopia	3,423	77,261	160,014	237,275	0.5%
Haya	Tanzania	1,306	19,864	21,639	41,503	0.2%
Pare	Tanzania	1,000	15,001	18,819	33,820	0.1%
Chonyi	Kenya	370	1,359	1,291	2,650	0.1%
Swahili	Kenya	138	2,668	4,200	6,868	0.0%
Tigrinya	Ethiopia	18	408	90	498	0.0%

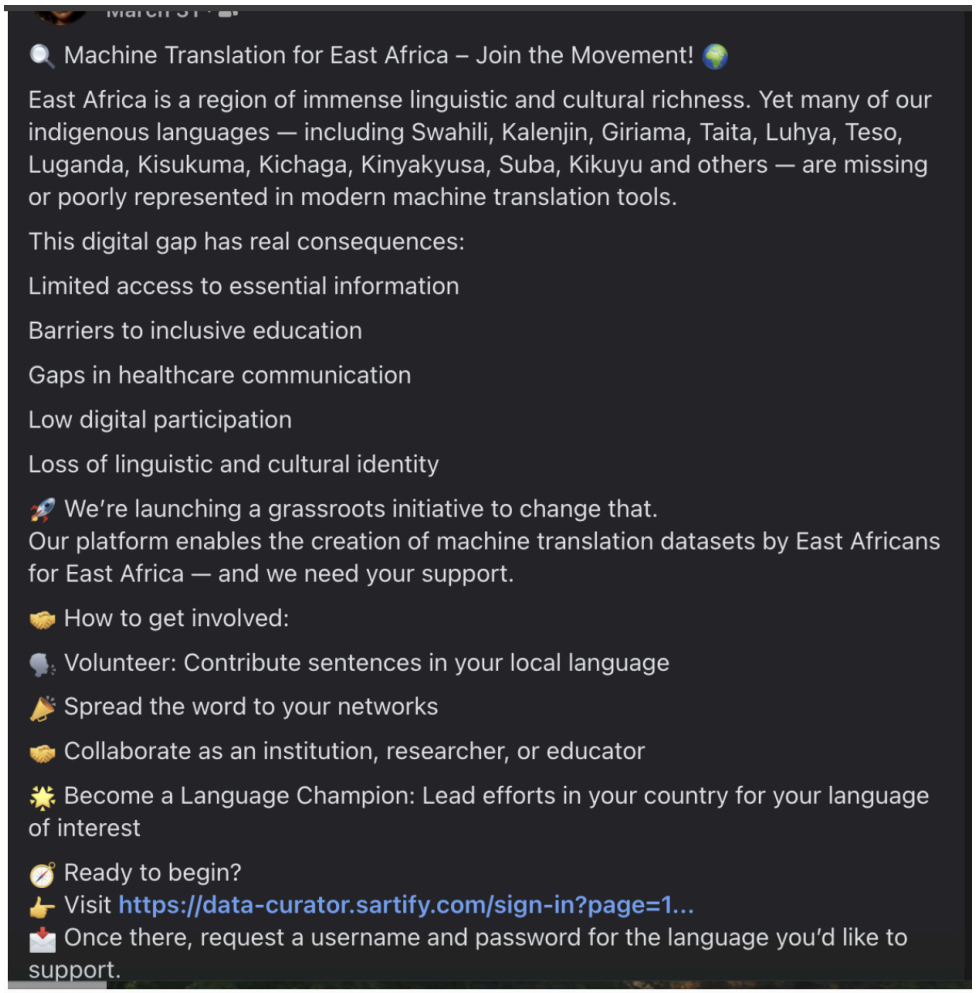

Figure 5: Community engagement post announcing the AfriMMT-EA initiative. The campaign invited volunteers, researchers, and institutions across East Africa to contribute translation data for low-resource languages, promoting inclusive participation and grassroots dataset creation.

Name	Country	Translation pairs	Old Tokens	New Tokens	Domain	source
Swahili	Kenya	138	2,668	-	Short stories, Politics, Healthcare	Publications
Duruma	Kenya	5,011	-	248,717	Religion	Bible
Suba	Kenya	7,965	-	437,423	Religion	Bible
Taita	Kenya	58,537 (new: 5,219)	802,080	86,177	Religion, Mixed domain (news, education, government, politics, healthcare)	Bible, Publications (Mbogho et al., 2025)
Meru	Kenya	5,420	-	206,282	Religion	Bible
Gusii	Kenya	6,494	-	261,580	Religion, Lexical, Mixed domain (news, education, government, politics, healthcare)	Bible, Publications (Mbogho et al., 2025)
Saamia	Kenya	107	-	289	Lexical	Web Articles
Kamba	Kenya	8,210	-	220,388	Lexical, Religion, Mixed domain	Bible, Web Articles
Teso	Kenya	899	-	1,467	Lexical (Wordlist)	Web Articles (Internet Archive)
Kalenjin	Kenya	7,969	-	7,969	Lexical, Healthcare, Survey, Mixed domain (news, education, government, politics, healthcare)	Web Articles, Private Data
Embu	Kenya	5,349	-	226,948	Religion	Bible
Chonyi	Kenya	370	-	1,291	Lexical	Web Articles
Luwanga	Kenya	7,337	-	7,736	Dictionary	Web Articles
Kikuyu	Kenya	100,195 (new pairs: 5,011)	3,419,801	90,569	Agriculture, Religion, Lexical	NLLBv1 dataset, OPUS, MT560, CLICS, Bible
Luo	Kenya	10,038	349,277	7,484	Children's Literature / African Literature, Mixed domain (news, education, government, politics, healthcare)	Children's Literature / African Literature
Maasai	Kenya	4,375	-	61,863	Religion, African literature, Lexical, Mixed domain (news, education, government, politics, healthcare)	Bible, Translation of Existing Publications (Mbogho et al., 2025), Web Articles
Borana	Kenya	9,697	-	264,151	Religion	Bible, Dictionary
Tharaka	Kenya	5,249	-	92,865	Religion	Bible
Bukusu	Kenya	4,240	-	22,311	Dictionary, Religious	Dictionary, Religion
Maragoli	Kenya	9,146 (new: 5,678)	23,445	35,632	Cultural Narrative (Folkloric and Oral Tradition), Conversational Folklore	Publications - KenCorpus, Web Articles
Chuka	Kenya	5,030	-	195,227	Religion	Bible
Samburu	Kenya	5,411	-	45,595	Dictionary	Dictionary
Marakwet	Kenya	5,249	-	266,270	Religion	Bible
Taveta	Kenya	5,249	-	186,804	Religion	Bible
Giriama	Kenya	9,629	-	376,830	Religion, Cultural Narrative, Oral Literature, Cultural Heritage, Mixed domain (news, education, government, law, politics, healthcare, political, public health)	Bible, Translation of Existing Publications (KenCorpus), Web Articles, Book Translation
Pokomo	Kenya	5,010	-	246,456	Religious	Bible
Pokot	Kenya	6,533	-	217,501	Religious	Bible
Somali	Kenya	5,011	-	235,446	Religious	Bible
Oromo	Ethiopia	3,423	160,014	-	Mixed	Published resources
Tigrinya	Ethiopia	18	90	-	Mixed	Published resources
Kinyarwanda	Rwanda	6,246	235,655	-	Religion	Bible
Bena	Tanzania	8,674	480,999	-	Religion, Mixed Domains	Bible and Published resources
Gogo	Tanzania	32,371	1,865,915	-	Religion	Bible
Makonde	Tanzania	9,111	516,338	-	Religion	Bible
Langi	Tanzania	7,849	409,292	-	Religion	Bible
Maasai	Tanzania	5,327	190,289	-	Religion	Bible
Jita	Tanzania	8,820	361,195	-	Religion, Mixed Domains	Bible and Published resources
Kihehe	Tanzania	7,849	365,460	-	Religion, Mixed Domains	Bible and Published resources
Kichaga	Tanzania	10,106	393,654	-	Religion, Mixed Domains	Bible and Published resources
Kisukuma	Tanzania	6,394	227,498	-	Literature, Mixed Domains	Published resources
Kihaya	Tanzania	1,306	21,639	-	Religion, Mixed Domains	Bible and Published resources
Pare	Tanzania	1,000	18,819	-	Mixed Domains	Bible
Acholi	Uganda	6,246	231,740	-	Religion, Mixed Domains	Bible and Published resources
Alur	Uganda	6,246	231,740	-	Religion	Bible
Aringa	Uganda	6,243	231,605	-	Religion	Bible
Ateso	Uganda	6,246	231,740	-	Religion	Bible
Gwere	Uganda	6,246	231,740	-	Religion	Bible
Jopadhola	Uganda	6,246	231,740	-	Religion	Bible
Kakwa	Uganda	6,246	231,740	-	Religion	Bible
Kebu	Uganda	6,246	231,740	-	Religion	Bible
Kumam	Uganda	6,246	231,740	-	Religion	Bible
Lango	Uganda	6,246	231,741	-	Religion	Bible
Lugbara	Uganda	162,446	3,079,301	-	Religion, Mixed Domains	Bible and Published resources
Masaaba	Uganda	6,246	231,741	-	Religion, Mixed Domains	Bible and Published resources
Nyole	Uganda	6,246	231,740	-	Religion	Bible
Nyoro	Uganda	6,246	231,740	-	Religion	Bible
Soga	Uganda	6,246	231,740	-	Religion	Bible

Table 4: Specific token and translation pair contributions for the 57 East African Languages considered

Pair Type	Lang Pair	n_samples	BLEU_270M	chrF++_270M	ter_270M	BLEU_1B	chrF++_1B	ter_1B
English and Swahili Paired	english→borana	889	0.1	3.36	103.98	0.31	3.86	104.61
	english→embu	20	0	0.65	100	0	0.65	100
	english→giriama	26	0.59	17.99	164.76	1.79	16.19	161.27
	english→gusii	20	1.28	2.49	104	0.64	3.96	104
	english→jita	21	0.46	2.33	102.56	0.85	5.96	105.13
	english→kamba	180	0.01	1.76	101.22	0	2.1	99.59
	english→kikuyu	18989	17.14	40.56	81.97	23.4	46.43	70.32
	english→luganda	23875	23.69	48.6	77.47	27.01	50.34	71.48
	english→luhya	1518	0	0.63	100.12	0	0.77	100.06
	english→luo	154	3.35	23.42	129.76	1.09	12.37	97.85
	english→maasai	416	8.95	29.11	106.15	9.11	30.04	104
	english→meru	39	1.38	7.53	107.46	0.45	8.68	107.46
	english→pokot	260	0.22	6.19	123.4	0.01	3.8	101.13
	english→swahili	26	0.4	13.02	203.32	0.21	12.59	216.34
	swahili→borana	1044	6.19	25.43	114.37	5.39	24.16	129.46
	swahili→embu	1044	6.19	29.84	114.47	9.71	33.69	94.98
	swahili→giriama	1885	20.94	47.6	77.75	11.99	31.1	90.17
	swahili→gusii	1269	4.64	31.43	125.3	4.92	26.75	129.29
	swahili→jita	1742	8.51	32.43	106.28	11.02	34	102.47
	swahili→kamba	1033	8.31	30.19	108.6	10.35	31.84	97.46
	swahili→kikuyu	1002	9.46	31.75	99.03	11.37	34.43	92.96
	swahili→luganda	1248	7.28	30.64	107.87	8.49	31.11	106.14
	swahili→luhya	2164	4.83	24.66	123.65	1.24	12.35	103.81
	swahili→luo	1577	6.97	32.93	99.26	2.18	13.58	98
	swahili→maasai	643	4.57	26.62	126.5	3.68	17.78	139.24
	swahili→meru	1044	2.86	21.47	150.08	4.64	20.57	112.11
	swahili→pokot	1044	2.86	23.21	129.8	2.11	18.1	140.49
	swahili→english	10844	23.77	50.19	67.53	1.27	17.41	87.62
English Paired Only	english→kalenjin	1431	8.22	29.55	127.9	12.17	35.17	105.29
	english→oromo	528	3.73	23.97	127.24	5.1	22.92	115
	english→samburu	1073	3.3	4.59	100.1	0.41	2.83	114.36
	english→tigrinya	4	12.13	25.56	150	9.29	7.42	200
Swahili Paired Only	swahili→acholi	1248	9.05	30.62	102.02	12.43	34.8	89.61
	swahili→alur	1249	11	34.94	94.73	11.85	36.27	92.52
	swahili→aringa	1247	5.27	15.46	103	4.14	11.08	112.1
	swahili→ateso	1249	5.11	27.39	116.06	6.56	28.4	105.55
	swahili→bena	1731	7.1	34.25	108.13	9.92	36.97	96.2
	swahili→chaga	1949	4.87	25.44	134.38	4.43	19.56	112.18
	swahili→chonyi	73	0	3.76	102.7	0	1.67	98.65
	swahili→chuka	951	7.23	29.23	112.43	8.9	30.85	103.49
	swahili→duruma	1001	7.37	32.21	108.46	6.53	26.5	102.94
	swahili→ganda	1247	11.53	34.51	93.53	16.44	40.11	88.07
	swahili→gogo	6451	6.64	26.05	114.3	7.16	26.76	94.93
	swahili→gwere	1248	3.37	28.27	122.51	4.95	30.71	104.61
	swahili→haya	247	2.96	30.71	94.88	0	5.58	98.08
	swahili→hehe	1570	5.32	32.36	115.94	4.75	29.58	119.42
	swahili→jopadhola	1249	7.37	28.49	99.41	8.35	29.91	95.35
	swahili→kakwa	1249	7.84	33.58	103.7	10.06	36.14	95.77
	swahili→kebu	1248	6.59	32.49	106.31	7.72	33.96	98.09
	swahili→kinyarwanda	1247	6.14	30.35	113.14	8.78	31.08	100.65
	swahili→kumam	1248	9.92	33.95	99.44	13.35	37.58	87.31
	swahili→langi	1568	9.32	32.78	99.6	10.64	33.85	94.96
	swahili→lango	1248	9.44	31.54	94.23	12.29	34.71	85.85
	swahili→makonde	1820	10.67	36.77	104.98	12.69	37.29	94.97
	swahili→marakwet	1044	5.17	25.56	115.81	4.93	24	117.04
	swahili→masaba	1249	5.73	31.15	105.25	8.15	34.46	97.58
	swahili→nyole	1249	2.81	25.87	143.33	4.01	27.55	111.88
	swahili→nyoro	1249	7.23	37.58	99.99	7.56	33.51	98.84
	swahili→pare	198	3.36	21.61	135.18	0	1.73	99.13
	swahili→pokomo	1001	5.46	30.74	122	6.83	28.84	101.42
	swahili→soga	1248	7.08	33.06	99.87	9.16	35.36	97.51
	swahili→somali	1000	4.22	28.66	103.67	5.17	29.42	99.53
	swahili→suba	1587	3.83	28.25	120.4	5.59	29.68	107.6
	swahili→sukuma	1277	13.68	44.52	93.45	4.41	20.31	100.4
	swahili→taita	5438	8.51	32.7	105.07	2.46	16.74	100.17
	swahili→taveta	1044	3.32	26.38	137.35	4.33	27.38	125.51
	swahili→zanaki	1561	6.54	34.59	101.81	7.55	35.41	99.35
Averaged Summary		1918.43	6.47	26.35	111.78	6.45	23.74	106.41

Table 5: chrF++, BLEU, and TER Scores for 67 East African Language Pairs across different Model Sizes